# Evolutionary-scale prediction of atomic level protein structure with a language model

*Year: 2022* | *Citations: 3385* | *Authors: Zeming Lin, Halil Akin, Roshan Rao*

---

## Abstract

Artificial intelligence has the potential to open insight into the structure of proteins at the scale of evolution. It has only recently been possible to extend protein structure prediction to two hundred million cataloged proteins. Characterizing the structures of the exponentially growing billions of protein sequences revealed by large scale gene sequencing experiments would necessitate a break-through in the speed of folding. Here we show that direct inference of structure from primary sequence using a large language model enables an order of magnitude speed-up in high resolution structure prediction. Leveraging the insight that language models learn evolutionary patterns across millions of sequences, we train models up to 15B parameters, the largest language model of proteins to date. As the language models are scaled they learn information that enables prediction of the three-dimensional structure of a protein at the resolution of individual atoms. This results in prediction that is up to 60x faster than state-of-the-art while maintaining resolution and accuracy. Building on this, we present the ESM Metage-nomic Atlas. This is the first large-scale structural characterization of metagenomic proteins, with more than 617 million structures. The atlas reveals more than 225 million high confidence predictions, including millions whose structures are novel in comparison with experimentally determined structures, giving an unprecedented view into the vast breadth and diversity of the structures of some of the least understood proteins on earth.